Application of the Moore-Penrose Inverse of a Data Matrix in Multiple Regression

Douglas M. Hawkins
Department of Applied Statistics
University of Minnesota
St. Paul, Minnesota

and

Dan Bradu
National Research Institute for Mathematical Sciences
CSIR
Pretoria, 0001 South Africa

Submitted by George P. H. Styan

ABSTRACT

Consider an adjoined $n \times p$ matrix $Z = (Y : X)$ relating to the regression of a dependent variable Y on a set of predictors X. It is shown that the Moore-Penrose inverse of Z contains a useful summary of information about multiple regressions between any column of Z and all other columns, as well as a set of case diagnostics that may be used to identify outliers and influential points. Z and the inverse are dual, so that Z is itself a diagnostic indicator of multiple regressions in the inverse. It is shown how the inverse may be used as a case diagnostic for both leverage and outlyingness, and also provides information about the dependence of subset regressions on particular cases.

INTRODUCTION

In the area of multiple regression of an independent variable y whose values are in an $n \times 1$ vector Y on the $n \times (p-1)$ matrix of predictors X, attention has been given in the literature to the "catcher matrix" defined by $(X^T X)^{-1} X^T$ [see, for example, Mosteller and Tukey (1977), Velleman and Welsch (1981)]. The usual interest in the catcher matrix lies in the fact that it

655 Avenue of the Americas, New York, NY 10010 0024-3795/90/$3.50

transforms the data Y into the least squares regression estimate b via the identity $b = (X^TX)^{-1}X^TY$, the first term of which is simply the catcher matrix. Thus if the catcher matrix is known, the regression coefficients can be computed without substantial further arithmetic. Consider instead the matrix obtained by adjoining the matrices of predictors and the dependent variable:

$$Z = (Y : X). \tag{1}$$

The matrix Z is of order $n \times p$. In what follows, we shall assume that Z is of full column rank p. This assumption is partly one of convenience in that the major properties derived carry over, with minor modification, to the case of rank-deficient data matrices Z.

We shall show that the Moore-Penrose inverse of this matrix Z is itself a useful regression diagnostic (particularly in conjunction with the data matrix), that it provides alternative ways of computing some standard measures associated with regression, and that it clarifies some aspects of the theory of latent root regression analysis (Hawkins 1973; Webster, Gunst, and Mason 1974; Eplett 1978). We shall in fact describe it as a *dual* of the original data set, the duality being in the sense both of quadratic programming and of the description in Hawkins (1973). This terminology is also consistent with that of Chapter 6 of Dempster (1969), where the column spaces of Z and $Z^{(*)}$ (defined below) are described as dual. We define the *star* of the $n \times p$ rank p matrix Z by

$$Z^{(*)} = Z(Z^TZ)^{-1}. \tag{2}$$

It should be noted that $Z^{(*)}$ is just the transpose of the Moore-Penrose inverse of Z. [For a general discussion of the use of generalized inverses in statistics, see the standard texts such as Pringle and Rayner (1971), Rao and Mitra (1971), or the article by Styan (1983).] We use this slightly nonstandard form in preference to the conventional Moore-Penrose inverse because it leads to Z and $Z^{(*)}$ having the same shape, both being $n \times p$ matrices. We shall use the abbreviation MPIT for this generalized inverse of a matrix, and term the elements and submatrices of $Z^{(*)}$ the *images* of the corresponding elements or submatrices of Z. The major concern of this paper is the exploration of the relationships between the properties of the variables in Z and those of the images of these variables in $Z^{(*)}$.

Some of the results to be discussed (notably the first four properties) are to be found in the thesis by Haber (1975). Another related piece of research is that of Guttman (1953), who writes Z as $Z = P + E$, where the elements of P are termed the *partial image scores* and consist of the predictions of each

column of Z from the remaining columns, and those of E are termed the *partial antiimage scores*. The matrix of partial antiimages, E, is just the matrix $Z^{(*)}$ with the columns differently scaled.

We should stress that most of the algebraic results given in this paper are, or follow easily from, standard theorems relating to the Moore-Penrose inverse. The contribution of this paper lies in the application of these properties to regression, rather than in the area of new theorems in algebra. Where proofs are given of properties, this is largely for completeness rather than indicating algebraic novelty. The proofs that are given are set out using general matrix results. We are grateful to the referee for pointing out that the results may be shown very elegantly using properties of projectors, an approach that we have not followed in deference to the needs of many users of regression.

NOTATION

We define the following additional notation:

$b = (X^TX)^{-1}X^TY$ is the least squares estimate of the regression coefficient vector;

$S = Y^TY - Y^TXb$ is the residual sum of squares from the regression;

$s = [S/(n - p + 1)]^{1/2}$ is the residual standard deviation;

$\bar{Z}$ is the $p \times 1$ column vector of means of Z.

It is very convenient to have a terse notation for certain submatrices of $Z^{(*)}$, and for this we shall also use the notation $Z^{(*)} = (Y^* : X^*)$, so that Y^* is the image of Y in $Z^{(*)}$ and X^* the image of X, and use $z_i = (y_i, x_i)$ and $z_i^* = (y_i^*, x_i^*)$ for the ith rows of Z and $Z^{(*)}$ respectively. This notation must be distinguished from the $^{(*)}$ notation: in particular X^* is not the same as $X^{(*)}$.

In regression, one almost invariably includes an intercept term. This may be done in either of two ways—(i) by including in X a column of 1's, and (ii) by precentering the data, subtracting off the mean from each column of Z, and fitting to the deviations from zero. In the latter case, the intercept term is recovered from the fitted coefficients and the mean vector. These methods are algebraically (though not computationally) equivalent. We shall make no notational distinction between the two approaches at this point, and the results given are equally applicable to either approach, and also to the no-intercept case. However, we shall make some comparisons between the two methods later.

Write the singular value decomposition (SVD) and QR decomposition of Z as

$$Z = P\Lambda V^T, \tag{3a}$$

$$Z = QR, \tag{3b}$$

where P, V, and Q are orthogonal matrices, and Λ is diagonal with positive elements. When Λ is limited to the p strictly positive diagonal elements, V is orthogonal, and P is column orthogonal. Conventionally R $(p \times p)$ is upper triangular, though for many purposes for which this decomposition is useful, other shapes for R are more helpful.

The first of the useful properties is the connection between these two decompositions of Z and the corresponding ones of $Z^{(*)}$.

PROPERTY 1. *The SVD and QR decompositions of* $Z^{(*)}$ *are*

$$Z^* = P\Lambda^{-1}V^T = Q(R^T)^{-1} = QR^{(*)}, \tag{4}$$

$R^{(*)}$ *being the* MPIT *of* R.

These two formulas are well known (see for example Kennedy and Gentle 1980, Dempster 1969) and are easily proved from first principles. An immediate implication is that the singular values of $Z^{(*)}$ are the inverses of those of Z, a property that will later be shown to be of practical statistical use.

The QR decomposition provides the preferred and numerically stable method of computing the MPIT via the inversion of R. While for expository purposes we chose to define $Z^{(*)}$ in terms of Z^TZ, computation based on one of these stable methods should be used, and not the form based on inversion of Z^TZ.

PROPERTY 2. *The matrices* Z *and* $Z^{(*)}$ *are dual—i.e.,* $Z = (Z^{(*)})^{(*)} = Z^{(*)}(Z^{(*)T}Z^{(*)})^{-1}$.

This immediate consequence of Property 1 implies that any property that $Z^{(*)}$ may have as a summary statistic of the regression of Y on X is exactly mirrored by the property that Z has as a summary of the regression of the image Y^* on X^*.

It follows at once from the definition of $Z^{(*)}$ that $(Z^TZ)^{-1} = (Z^{(*)T}Z^{(*)}) = (R^{(*)T}R^{(*)})$, so that $R^{(*)}$ is an inverse square root of Z^TZ, whose properties as a regression summary statistic are discussed in Hawkins and Eplett

(1982). This matrix in many of its possible forms is an important diagnostic of the various multiple regressions connecting the elements of Z.

PROPERTY 3. *The (ij)th element of $Z^{(*)}$ is the scaled residual from the ith case when the jth variable of Z is regressed on all other variables. The scaling factor is the inverse of the residual sum of squares of this regression.*

Proof. We shall give a direct demonstration of this result. For illustrative purposes, we shall prove it for Y^*, the image of Y in $Z^{(*)}$, the application to the other variables following in the same way. From the definition, we obtain

$$Z^{(*)} = (Y : X)\begin{pmatrix} Y^TY & Y^TX \\ X^TY & X^TX \end{pmatrix}^{-1}.$$

Application of the inversion theorem for partitioned matrices shows at once that the ith row of this matrix (corresponding to the ith case in the data set) is

$$\frac{y_i - x_i b}{S}, \quad x_i(X^TX)^{-1} - \frac{(y_i - x_i b)b^T}{S}, \tag{5}$$

the required result. As no special use was made of the fact that in the original problem Y was regarded as different from the columns of X, it follows that the result is true for all columns of $Z^{(*)}$. ■

Clearly, $\Sigma_i(y_i^*)^2 = S/S^2 = 1/S$, and so the squared Euclidean norm of Y^* is the inverse of the residual sum of squares of the regression. The column as a whole therefore contains both the residual sum of squares of the regression and the individual residuals.

This property carries over to the predictors as well. Here, it should be recalled that the images of the predictors in X are based on regressions using both the other predictors and the dependent variable, and therefore differ from partial residuals as discussed for example in Belsley, Kuh, and Welsch (1980).

Going a stage further, the sum of squares and cross products matrix of $Z^{(*)}$ is also informative:

PROPERTY 4. *The ijth element of $W = Z^{(*)T}Z^{(*)}$ is proportional to the regression coefficient of the jth column when the ith column of Z is regressed on all other columns. The constant of proportionality, given by the*

diagonal of W, is the residual sum of squares of the regression — specifically (illustrating with $i = 1$),

$$W = \begin{pmatrix} 1/S & -b^T/S \\ -b/S & (X^TX)^{-1} + bb^T/S \end{pmatrix}.$$

For centered data, this property is very familiar, since W is the inverse of the covariance matrix of $(Y: X)$, and its proof is an easy consequence of the formula for the inverse of a partitioned matrix. From the statistical point of view, it can be interpreted as saying that the covariances between the elements of $Z^{(*)}$ provide the multiple regression coefficients between the elements of Z and (by duality) vice versa. An alternative viewpoint is obtained by rewording Guttman's Theorem 3—the correlation between a pair of columns of $Z^{(*)}$ is the negative of the partial correlation between the same pair of columns of Z adjusted for the remaining $p - 2$ columns.

The definitions immediately give another result linking the primal and dual data sets:

PROPERTY 5. $ZZ^{(*)T} = Z^{(*)T}Z = Z(Z^TZ)^{-1}Z^T$.

This matrix is the projector onto the column space of Z, or equivalently (since it has the same column space), of $Z^{(*)}$. In regression, this matrix has been suggested by Gray and Ling (1984) as a possible diagnostic for the detection of subsets of jointly "influential" cases—cases which are of high leverage, outlying, or both. These authors suggest that a cluster analysis of its elements can find such influential sets of cases.

The Mahalanobis distance of the ith case from the origin, using as the distance metric the unnormalized sum of squares and cross products of the z_i, is an element of this matrix—it is

$$D_{zi}^2 = z_i z_i^{*T} = z_i (Z^TZ)^{-1} z_i^T. \tag{6}$$

The work of Gnanadesikan and Kettenring (1972) on multivariate outliers distinguishes between two types of outlier. Both types have large Mahalanobis distances D_{zi}^2, but while Type A outliers also have large Euclidean norms $\|z_i\|$, Type B outliers do not. Type A outliers are easier to locate by univariate methods, since their great Euclidean distance from the rest of the data implies that they must stand out from the other data in one or more univariate plots. Type B outliers however are much harder to find. They

stand out only on the minor components, and are therefore undetectable using univariate plots.

Using Property 5 and the Cauchy-Schwartz inequality however, we see immediately that D_{zi}^2 is bounded by the product of the Euclidean distances of z_i and z_i^* to the origin—

$$D_{zi}^2 \leqslant \|z_i\| \, \|z_i^*\|.$$

This shows that Type A outliers in the primal data set are Type B outliers in the dual, and conversely, since if the Mahalanobis distance D_{zi}^2 is large while $\|z_i\|$ is small, then $\|z_i^*\|$ must be large. This means that univariate study of the variables in $Z^{(*)}$ is a valuable supplement to that of the variables of Z, in that the one locates Type A aberration, while the other locates Type B aberration.

We now turn to the issue of the affect of centering the data Z as against a column of 1's in the matrix X for fitting the intercept, and the effect of these two options on the MPIT. This issue is easily resolved for the nontrivial columns. The residuals and the residual variance are unaffected by the computational question of whether the regression is fitted with the intercept term explicitly included as a predictor, or by precentering the data, fitting the regression on centered data, and then inferring the intercept term. Since the columns of $Z^{(*)}$ are the residuals of the corresponding variables in the regression, scaled by the residual variance, the images of Y and the nontrivial columns of X are the same in the precentered form of Z as in the extensive form in which a column of 1's is included in Z. The two forms do however differ in that in the extensive form $Z^{(*)}$ has an additional column, 1* say, that is the image of the column of 1's in Z, which one might suppose to contain potentially useful information. Writing the expression for 1_i^*, the ith element of the column 1*, explicitly in terms of the image z_i^* obtained after precentering all columns of Z, we get

$$1_i^* = n^{-1} - \bar{Z} z_i^{*T},$$

and so, apart from the additive constant n^{-1}, it is simply a sum of the scaled residuals of the different variables weighted by the corresponding elements of $\bar{Z}$. We shall see in the example that the 1* column provides a diagnostic of multicollinearity and the extent to which this is masked by individual cases, and so we consider the 1* column generally worth computing and reporting, even if we prefer to carry out the computations using precentered data.

COMPUTATION OF CONVENTIONAL REGRESSION DIAGNOSTICS

An excellent coverage of the common regression diagnostics for cases may be found in Belsley, Kuh, and Welsch (1980). The major indicators are various functions of

(i) the residual from the regression $e_i = (y_i - x_i b)$;
(ii) the standardized residual e_i/s;
(iii) the leverage or potential of the predictors;
(iv) the Studentized residual $e_i/s\sqrt{1 - D_{xi}^2}$.

All these quantities can be recovered from the MPIT with the original data matrix.

The residual standard deviation is given by

$$s = \frac{1}{\sqrt{(n-p)\Sigma y_i^{*2}}}.$$

As the individual elements of the image of the dependent variable are just the scaled residuals, the original and the standardized residuals may be recovered simply by undoing the scaling:

$$e_i = y_i^* \mathrm{S}, \qquad e_i/s = y_i^* (\mathrm{S}/s).$$

For comparison of the residuals of different cases, it is not necessary to transform to the original or standardized residuals: the y_i^* may be compared directly.

Next, the leverage is commonly measured by the Mahalanobis distance of the case to either the origin (if the model has an explicit intercept term) or the mean $\bar{x}$ (if the data are precentered). In either case, denoting the distance by D_{xi}^2, we may compute it by using the result

$$D_{zi}^2 = D_{xi}^2 + \left(\frac{y_i - x_i b}{\mathrm{S}}\right)^2 = D_{xi}^2 + y_i^{*2}\mathrm{S},$$

which shows that by subtracting the squared scaled residual $y_i^{*2}\mathrm{S}$ from D_{zi}^2, we obtain D_{xi}^2. From these two measures— D_{xi}^2 and the scaled residual—the Studentized residuals of the cases and the whole range of conventional diagnostics can be computed: (i) D_{xi}^2 measures the leverage of the point, (ii)

the residual and the standardized residual measure the fit of the point to the line, and (iii) the Studentized residual is the likelihood ratio test for the hypothesis that the dependent variable of the case differs from what might be inferred from the other points. Thus all three have a place in the overall picture. In addition to these standard measures, the distance D_{zi}^2 (the diagonal of the Gray-Ling matrix) is a useful summary statistic, high values indicating that the case is of high leverage, outlying, or both.

UPDATING AND DOWNDATING

It is common that a sufficiently outlying or a sufficiently leveraged case can so obscure the overall structure of the data as to mask some other deficiencies or structure. For this reason, it will usually be advisable to carry out several analyses of the data, each consisting of a study of the diagnostic statistics following either addition or deletion of a case. These considerations then lead to the question of the method of modifying the MPIT following updating (adding) or downdating (deleting) a case. Similarly, for purposes of studying subset regressions, there may be interest in the addition or deletion of variables, and formulas are required for this operation as well. Formulas for both these updates and downdates are given in a unified form in Graybill (1969). We repeat these formulas in our notation for completeness.

Updating by Case

Let z be the new case, and $Z^{(*)}$ the "old" MPIT. Define $c = z(Z^{(*)T}Z^{(*)})$ and $d = cz^T$. Then the updated MPIT is

$$\begin{pmatrix} Z^{(*)} - z^T c/(1+d) \\ c/(1+d) \end{pmatrix},$$

and the updated $(Z^TZ + zz^T)^{-1}$ is given by

$$Z^{(*)T}Z^{(*)} - \frac{c^Tc}{1+d},$$

and so the MPIT is easily updated to include a new case.

To downdate, writing z for the case to be deleted, the downdated $Z^{(*)}$ and $Z^{(*)T}Z^{(*)}$ are given by simply reversing the plus and minus signs in the above formula. This formulation retains a row in $Z^{(*)}$ for the deleted case

which is very useful, as it contains, in place of the usual scaled residuals, predicted residuals (Allen 1974) for the deleted case on the regressions fitted to the retained cases.

If several cases are to be deleted, the same downdating procedure may be used repeatedly, with flags being set on the deleted cases. In this way, the predicted residuals are obtained for the successive regression models obtained for the different subsets of cases that appear to be of interest.

Updating and Downdating by Variable

Adding a variable consists of adjusting the MPIT to include the image column of a variable not previously there. Let the original matrix Z be updated by the addition of a new column, w say. Then the MPIT of the partitioned matrix $(Z: w)$ is obtained by first projecting w onto the null space of Z^T by forming

$$f = (I - Z^{(*)}Z)w \quad \text{and} \quad d = \frac{f}{\|f\|^2},$$

and the updated MPIT is

$$(Z: w)^{(*)} = (Z^{(*)} - dw^T Z^{(*)}: d).$$

The rightmost column obviously accords with the characterization of the MPIT as having columns which are the inverse-variance-scaled residuals of that variable on all others.

Downdating by variable consists of removing a variable from the model. We shall illustrate this process by the removal of Y from the model. Recall that in partitioned form, $Z^{(*)}$ is

$$\begin{pmatrix} \dfrac{Y - Xb}{S} & X(X^TX)^{-1} - \dfrac{(Y - Xb)b^T}{S} \end{pmatrix},$$

while

$$Z^{*T}Z^{(*)} = \begin{pmatrix} 1/S & -b^T/S \\ -b/S & (X^TX)^{-1} - bb^T/S \end{pmatrix}.$$

Thus the updating consists of computing S and b from $Z^{*T}Z^*$, and adding back the subtrahend in the partitioned form of Z^*. This then gives $X(X^TX)^{-1}$ as the right component of the partition. The scaled residuals of Y^* can then conveniently be replaced by the original Y, leaving the MPIT ready for subsequent updating if desired.

Numerical stability in these up- and downdates is obtained by using, not these Z^TZ-based computations, but the formulas for the up- and downdating of the QR decomposition by cases and variables which may be found in Gragg, LeVeque, and Trangenstein (1979).

THE USE OF THE MPIT IN SUBSET AND LATENT ROOT REGRESSION

Next, we explore some properties of the conjunction of the MPIT and the original data matrix Z in elucidating some aspects of subset regression, as, in particular, explored using latent root regression (Hawkins 1973; Webster, Gunst, and Mason 1974). While retaining no essential distinction between predictor and dependent variables, we shall continue to suppose that the first of the variables is a dependent, and the remainder are predictor variables. Consider the adjoined data matrix $(Z: Z^*)$. As Z and Z^* form a bilinear pair of bases for their column space, the sum of squares and cross products matrix of the adjoined matrix is

$$(Z: Z^*)^T(Z: Z^*) = \begin{pmatrix} Z^TZ & I \\ I & (Z^TZ)^{-1} \end{pmatrix}.$$

This form has a number of interesting consequences, not all of which will be explored here. The first interesting result connects the regression of Y on X with Y^* on X^*. This is:

PROPERTY 6. *The proportion of variance explained by the multiple regression of Y on X is identical to that of Y^* on X^*.*

Proof. Write $(Z^TZ)_{ij}$ for the ijth element of Z^TZ, and similarly for $Z^{(*)T}Z^{(*)}$. Then from earlier results, the proportion of variance explained by the regression of Y on X is $1 - 1/[(Z^TZ)_{11}(Z^{(*)T}Z^{(*)})_{11}]$. However, from the duality, this is also exactly equal to the proportion of the variance of Y^* explained by the regression of Y^* on X^*. ■

Let us now consider the subset regression of Y on only some of the components of X. To slightly simplify discussion, suppose that all columns of Z have been centered, partition X as $(X_1 : X_2)$, and partition $X^{(*)}$ conformably as $(X_1^* : X_2^*)$. We shall show that the regression of Y on X_1 omitting X_2 is equivalent in many respects to the regression of Y^* on X_2^* omitting X_1^*.

PROPERTY 7. *Let X_2^* be partialed out of the dual regression, let $c_{1,2}^*$ be the vector of partial covariances between Y^* and X_1^* given X_2^*, and let S_2^* be the residual sum of squares of Y^* on X_2^*. Then the regression coefficient vector of Y on X_1, b_1 say, is given by*

$$b_1 = -c_{1.2}^*/S_2^*.$$

PROPERTY 8. *Let $e_{.1}$ denote the vector of residuals of Y on X_1 omitting X_2, and $e_{.2}^*$ that of residuals of Y^* on X_2^* omitting X_1^*. Let s_1 and s_2^* denote the corresponding residual standard deviations. Then*

$$e_{.1}/s_1 = e_{.2}^*/s_2^*,$$

so that the scaled residuals of the primal regression omitting X_2 are identical to those of the dual regression on X_2^*.

We shall show both these results by looking at something rather more general. Write Z and $Z^{(*)}$ in conformable partitioned form as $Z = (Z_1 : Z_2)$, $Z^{(*)} = (Z_1^* : Z_2^*)$. Then writing $Z^{(*)T}Z^{(*)} = W$ also in conformable partitioned form, partialing Z_2^* out of Z_1^* consists of replacing Z_1^* by

$$Z_{1.2}^* = Z_1^* - Z_2^* W_{22}^{-1} W_{21} = (Z_1^* : Z_2^*) \begin{bmatrix} I \\ -W_{22}^{-1}W_{21} \end{bmatrix}$$

$$= (Z_1 : Z_2) \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} I \\ -W_{22}^{-1}W_{21} \end{bmatrix} = Z \begin{bmatrix} W_{11} - W_{12}W_{22}^{-1}W_{21} \\ 0 \end{bmatrix},$$

which we immediately recognize as $Z_1^{(*)}$— the MPIT of the submatrix Z_1.

This means that the operation of partialing Z_2^* out of Z_1^* is no more than a downdate of $Z^{(*)}$ to remove the variables making up the columns of Z_2. Property 7 is then easily recognized as simply Property 4 applied to the problem with the variables in Z_2 removed. An interesting discussion of Property 7 may also be found on p. 113 of Dempster (1969), where the results of partial sweeping of Z^TZ and of $Z^{(*)T}Z^{(*)}$ are discussed.

Property 8 may be inferred at once by applying Property 3 to Z_1. This tells us that the regression residuals of Z on X_1 are the values of $Z^*_{1.2}$ rescaled by the residual variance.

Properties 7 and 8 explain the observations in Hawkins (1973) and Webster et al. (1974). Reinterpreted in the terms of this paper, while neither of these papers explicitly identified the dual variables, their authors show that the elimination of a block of predictors in the primal problem is equivalent in several of its properties to the inclusion of the images of those same variables in a regression of the dual image as dependent variable.

These properties of $Z^{(*)}$ show that it provides a powerful set of graphic diagnostics. Consider the plot of Y^* against one of the columns of X^*. We will term such a plot an *image plot*. Since the strength of linear association in the image plot provides a direct measure and test for the importance of the corresponding primal variable in the multiple regression, the plot provides a case diagnosis indicating which cases contribute to, and perhaps which cases mask, the contribution of that primal variable. From this description and the fact that the correlation coefficient of the image plot is identical (with sign reversed) to the corresponding added variable plot, we might suppose that the plot is no more than a standard added variable plot, but this is not the case at all. Points which are Type A outliers in the added variable plot will be Type B outliers in the image plot, and conversely, leading to different interpretations of the cases.

EXAMPLE

We shall illustrate some of the diagnostic uses of the MPIT by a consideration of the data set created by Hocking and Pendleton (1983). This data set was deliberately set up to confound univariate diagnostics. It contains 22 "good" points that are nearly collinear; one high leverage point obscuring the collinearity; and three outliers, one of which is also of high leverage. The primal data set has been rescaled so that Z^TZ has 1's on the diagonal. This absolute scaling facilitates comparison of the different variables in Z. The resultant $(Z: Z^{(*)})$ matrix is shown in Table 1. Along with each case the derived quantities $\|z_i\|$, $\|z_i^*\|$, D^2_{zi}, and D^2_{xi} are also given.

As intended by Hocking and Pendleton, there are no really striking features in either Z or the set $\|z_i\|$. Turning to the dual variables, however, the structure of the data set is uncovered at once. The images of the dependent variable in cases 11, 17, and 18 are huge relative to those of the other cases, which immediately flags these three observations as possible outliers.

TABLE 1

ORIGINAL DATA FOR Z, $Z^{(*)}$ AND CASE DIAGNOSTICS

i	z_i					z_i^*					Plot symbol	Stud. res.	$\|z_i\|$	$\|z_i^*\|$	D_{zi}^2	D_{xi}^2
1	0.209	0.170	0.010	0.382	0.196	1.78	−4.31	−.20	0.04	2.86	A	0.31	0.51	5.47	0.218	0.215
2	0.215	0.187	0.064	0.259	0.196	4.88	−5.66	0.46	−.17	0.72	B	0.78	0.44	7.52	0.115	0.093
3	0.203	0.203	0.167	0.113	0.196	1.47	−.70	0.17	−.13	−.61	C	0.23	0.40	1.75	0.050	0.048
4	0.202	0.198	0.149	0.160	0.196	0.18	0.31	−.06	0.02	−.25	D	0.03	0.41	0.43	0.042	0.042
5	0.187	0.200	0.221	0.078	0.196	−.98	1.07	−.06	.07	0.22	E	−.15	0.41	1.47	0.053	0.052
6	0.219	0.224	0.188	−.002	0.196	3.93	−.78	0.56	−.32	−3.17	F	0.65	0.41	5.15	0.169	0.155
7	0.220	0.202	0.086	0.178	0.196	1.51	−.50	−.16	−.11	−.59	G	0.24	0.41	1.70	0.084	0.081
8	0.136	0.167	0.335	0.116	0.196	0.52	−5.54	0.81	−.17	4.62	H	0.09	0.46	7.28	0.302	0.301
9	0.221	0.223	0.161	0.010	0.196	1.68	1.00	−.04	−.34	−2.17	I	0.28	0.40	2.94	0.157	0.155
10	0.200	0.172	0.064	0.334	0.196	2.29	−4.63	0.12	0.04	2.41	J	0.38	0.47	5.70	0.152	0.147
11	0.215	0.211	0.071	0.080	0.196	−12.87	12.98	−3.34	−.16	3.09	K	−2.16	0.37	18.83	0.328	0.174
12	0.196	0.187	0.128	0.212	0.196	−1.57	0.60	−.46	0.05	1.53	L	−.25	0.42	2.32	0.055	0.053
13	0.193	0.169	0.083	0.356	0.196	0.76	−3.02	0.05	0.25	2.22	M	0.13	0.49	3.83	0.164	0.163
14	0.152	0.186	0.327	0.040	0.196	−1.05	−1.69	0.36	−.19	2.75	N	−.18	0.45	3.42	0.175	0.174
15	0.229	0.199	0.038	0.235	0.196	3.13	−2.09	−.00	−.09	−.77	O	0.51	0.43	3.84	0.131	0.122

TABLE 1 *Continued*

i	z_i					z_i^*					Plot symbol	Stud. res.	$\|z_i\|$	$\|z_i^*\|$	D_{zi}^2	D_{xi}^2
16	0.166	0.206	0.334	−.065	0.196	−1.87	1.18	0.04	−.32	1.09	P	−.31	0.47	2.48	0.180	0.177
17	0.212	0.195	0.151	0.157	0.196	14.62	−14.53	2.63	−.51	−1.76	Q	2.28	0.41	20.85	0.240	0.041
18	0.181	0.185	0.099	0.323	0.196	−23.61	24.10	−4.35	1.28	2.45	R	−3.82	0.47	34.12	0.631	0.113
19	0.177	0.214	0.305	−.066	0.196	−4.66	5.49	−.59	−.16	−.00	S	−.78	0.46	7.22	0.180	0.160
20	0.225	0.215	0.123	0.082	0.196	4.39	−2.20	0.40	−.33	−2.07	T	0.71	0.40	5.35	0.132	0.114
21	0.165	0.177	0.240	0.147	0.196	1.73	−5.13	0.57	−.22	3.25	U	0.28	0.42	6.34	0.122	0.119
22	0.195	0.186	0.141	0.195	0.196	3.07	−4.66	0.41	−.20	1.58	V	0.48	0.41	5.82	0.064	0.055
23	0.202	0.207	0.181	0.080	0.196	−.22	1.27	−.14	−.12	−.64	W	−.04	0.40	1.45	0.059	0.059
24	0.205	0.216	0.269	0.343	0.196	2.22	9.28	2.46	2.26	−15.13	X	1.25	0.56	18.20	0.932	0.927
25	0.156	0.174	0.270	0.153	0.196	1.04	−4.48	0.67	−.08	3.10	Y	0.17	0.44	5.58	0.160	0.159
26	0.184	0.208	0.261	−.010	0.196	−2.36	2.63	−.32	−.22	0.37	Z	−.38	0.43	3.57	0.106	0.101

$Z^{(*)T}Z^{(*)}$

1075.4	−1060.4	203.0	−34.1	−162.8
−1060.4	1282.3	−185.0	59.5	−108.5
203.0	−185.0	46.3	−1.5	−56.7
−34.1	59.5	−1.5	7.8	−30.2
−162.8	−108.5	−56.7	−30.2	343.3

TABLE 2

RESULTS WITH CASE 24 DELETED

i	z_i^*					Stud. res.	$\|z_v\|$	$\|z_i^*\|$	D_{zi}^2	D_{xi}^2
1	0.34	−10.32	−1.80	−1.42	12.68	0.06	0.51	16.51	0.247	0.247
2	4.32	−7.98	−.15	−.74	4.51	0.66	0.44	10.16	0.119	0.103
3	2.49	3.56	1.30	0.91	−7.56	0.37	0.40	8.86	0.064	0.059
4	1.61	6.28	1.53	1.47	−9.98	0.24	0.41	12.09	0.070	0.068
5	−.01	5.13	1.01	0.91	−6.39	−.00	0.41	8.30	0.066	0.066
6	5.74	6.80	2.57	1.52	−15.53	0.92	0.41	18.14	0.215	0.186
7	1.77	0.60	0.13	0.16	−2.37	0.27	0.41	3.02	0.085	0.082
8	−.34	−9.13	−.15	−1.04	10.48	−.06	0.46	13.94	0.312	0.312
9	1.93	2.06	0.24	−.09	−3.89	0.30	0.40	4.81	0.158	0.155
10	1.77	−6.78	−.45	−.49	5.91	0.28	0.47	9.19	0.155	0.153
11	−18.71	−11.39	−9.81	−6.08	42.83	−3.80	0.37	49.47	0.798	0.493
12	−1.62	0.39	−.52	−.00	1.87	−.24	0.42	2.56	0.055	0.053
13	1.90	1.74	1.31	1.40	−5.55	0.30	0.49	6.41	0.181	0.178
14	−1.39	−3.12	−.02	−.54	5.09	−.22	0.45	6.15	0.177	0.175
15	3.47	−.66	0.38	0.26	−3.10	0.54	0.43	4.72	0.132	0.122
16	−2.37	−.89	−.51	−.83	4.46	−.38	0.47	5.22	0.184	0.179
17	16.08	−8.42	4.26	0.97	−11.72	2.38	0.41	22.04	0.269	0.044
18	−19.74	40.23	−.07	5.20	−23.86	−4.03	0.47	51.03	0.839	0.499
19	−4.15	7.61	−.03	0.35	−3.47	−.66	0.46	9.34	0.184	0.169
20	4.87	−.19	0.94	0.16	−5.35	0.75	0.40	7.30	0.135	0.114
21	0.51	−10.26	−.79	−1.47	11.61	0.08	0.42	15.58	0.143	0.143
22	2.23	−8.17	−.53	−1.05	7.31	0.33	0.41	11.25	0.074	0.070
23	0.63	4.81	0.80	0.74	−6.41	0.09	0.40	8.11	0.069	0.068
24	32.49	135.61	36.01	32.97	−221.18	1.27	0.56	265.98	13.614	12.695
25	1.22	−3.72	0.87	0.11	1.86	0.19	0.44	4.42	0.160	0.159
26	−2.56	1.82	−.54	−.42	1.70	−.39	0.43	3.63	0.107	0.101

$Z^{(*)T}Z^{(*)}$

1147.7	−758.9	283.1	39.2	−654.5
−758.9	2540.8	149.2	365.4	−2160.9
283.1	149.2	135.1	79.7	−601.7
39.2	365.4	79.7	82.2	−529.1
−654.5	−2160.9	−601.7	−529.1	3690.7

TABLE 3

AFTER DELETION OF CASES 11, 17, 18, AND 24

i	z_i^*				Stud.	$\|z_i\|$	$\|z_i^*\|$	D_{zi}^2	D_{xi}^2
1	− 5	− 14	− 6	58	− 1.46	0.51	70.0	0.392	0.323
2	− 45	2	− 4	15	1.13	0.44	57.0	0.205	0.153
3	9	6	4	− 31	0.50	0.40	36.8	0.102	0.092
4	41	5	8	− 47	− .18	0.41	64.0	0.189	0.188
5	33	2	5	− 26	− .41	0.41	44.3	0.121	0.113
6	− 20	18	5	− 58	2.32	0.41	88.0	0.423	0.224
7	6	− 4	− 1	10	− .44	0.41	17.8	0.110	0.101
8	− 31	6	0	− 2	1.19	0.46	43.0	0.364	0.319
9	− 19	− 3	− 4	23	0.09	0.40	30.9	0.219	0.219
10	− 2	− 2	0	10	− .20	0.47	12.5	0.166	0.164
11	63	− 108	− 42	363	− 5.20	0.37	492.2	8.275	3.061
12	41	− 11	0	18	− 1.76	0.42	67.7	0.204	0.067
13	53	8	11	− 61	− .29	0.49	82.4	0.430	0.428
14	− 6	0	0	3	0.10	0.45	8.1	0.178	0.178
15	− 3	0	0	0	0.13	0.43	4.8	0.148	0.147
16	− 9	− 7	− 5	35	− .60	0.47	40.8	0.237	0.224
17	151	56	7	− 122	7.12	0.41	299.6	2.865	0.134
18	431	− 56	29	− 47	− 6.29	0.47	564.5	9.903	2.792
19	51	− 12	0	17	− 2.16	0.46	79.4	0.355	0.171
20	− 40	6	− 2	− 3	1.49	0.40	57.0	0.244	0.155
21	− 44	0	− 5	29	0.71	0.42	56.5	0.212	0.193
22	− 40	0	− 5	27	0.65	0.41	51.9	0.143	0.126
23	24	0	2	− 13	− .43	0.40	30.0	0.094	0.086
24	624	257	200	− 1334	1.37	0.56	1549.1	83.611	76.660
25	− 7	11	4	− 36	1.17	0.44	49.4	0.254	0.202
26	15	− 11	− 4	35	− 1.39	0.43	55.6	0.205	0.125

$Z^{(*)T}Z^{(*)}$

17748	− 13327	4390	347	− 8457
− 13327	20322	− 1340	1967	− 7164
4390	− 1340	1468	515	− 4670
347	1967	515	495	− 3096
− 8457	− 7164	− 4670	− 3096	21761

This information could have been obtained from the conventional case diagnostics of the case leverage and Studentized residuals. The MPIT provides much additional information that is not obtainable from these usual summary case diagnostics, however—for example, all three suspicious cases have large images on the x_2, and so if they are outlying, it could as easily be due to an inconsistency in x_2 as to one in the dependent variable, illustrating the general fact that outliers can be in either the carriers or the dependent variable. It is a strength of the diagnostic use of the MPIT that it makes no real distinction between the dependent and the predictor variables, a consequence of which is the ability to show so clearly the possible alternative of an outlier in one carrier rather than in the dependent variable.

Comparative study of the D_{zi}^2 and D_{xi}^2 columns is also very informative. Concentrating on these three cases shows that they are not of particularly high leverage in the x space, although they are much more so in the z space. Thus they are low-leverage outliers.

The enormous leverage of case 24, clearly shown in the summary statistic D_{xi}^2, is explained by its 1* entry of -15.134, which dominates the entire

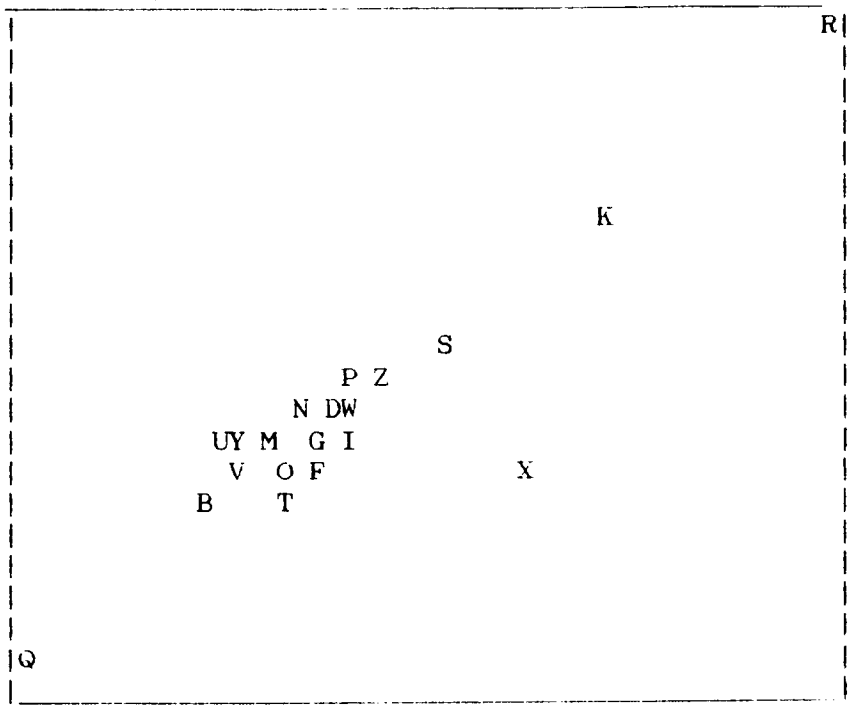

FIG. 1. Scatterplot of Y^* against x_1^*.

column and shows that there is a near-multicollinearity in the other cases that is not present in case 24.

While this single use of the MPIT has uncovered all the features built into the data set, one will not know this in practice. The presence in the sample of a point whose D_{xi}^2 and D_{zi}^2 are so dominant must create some worry about whether one has found the important features from the single pass of the MPIT, and to allay these worries, one would then carry out a downdate, deleting case 24. This gives the MPIT of Table 2. While images for case 24 remain in the table, they are scaled predicted residuals—images in terms of the metric of the retained cases. Table 2 confirms and reinforces the diagnosis of Table 1 as regards the outlyingness of cases 11, 17, and 18 and high leverage of case 24. In addition, though, the image vectors of the three nontrivial predictors are much longer than in Table 1, providing a clearer diagnosis of the nonpredictive multicollinearity between the three predictors which case 24 masked.

The relative uniformity of the D_{xi}^2 values does not support a particular need to strip off further points from the data set to get rid of the effects of

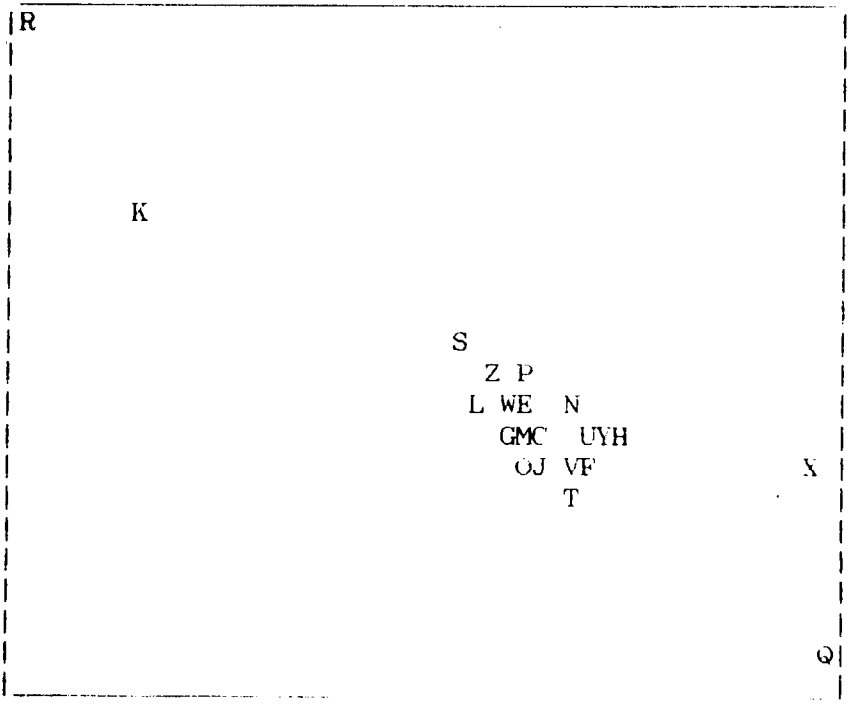

FIG. 2. Scatterplot of Y^* against x_2^*.

differing leverage, and one could terminate the analysis at this stage or perhaps (more for the sake of completeness than in the hope of making significant further discoveries) downdate the three outliers and see the effect on the MPIT. We have performed, but will not expand upon, this confirmatory analysis (Table 3).

A data set with only four predictors would not seem to be a particularly fertile one for studying subset methods, and this use of the MPIT will not be illustrated at any great length. The statistical significance and coefficient of determination of various subsets of the data are as follows:

Subset	t value for intercept	x_1	x_2	x_3	R^2
All data	1.33	9.66	−10.11	1.80	0.9358
Omitting 24	1.69	2.11	−4.84	0.75	0.9404
Omitting 11, 17, 18	4.33	37.90	−37.78	7.92	0.9959
Omitting 11, 17, 18, 24	1.67	4.37	−6.59	−0.18	0.9962

Of great interest is the situation as regards Y and the third column of x_3. In creating the data set, Hocking and Pendleton generated 22 "clean" data points, in which Y was a linear function of x_1 and x_2. To this set they then added x_3, a linear function of x_1 and x_2 with noise; the three outliers; and case 24, which deviated from the relationship between x_3 and (x_1, x_2). Thus in the clean set of points, there is no partial association between Y and x_3 and the correct regression of Y on x_1 and x_2 only. Any significance seen for X_3 is thus entirely due to case 24, and is masked by the outlying cases 11, 17, and 18.

Scatterplots of Y^* against the columns of X^*, as well as the more conventional added variable plots, are shown in Figures 1–4. In all the plots, the ith point is represented by the ith letter of the alphabet. The "problem points" are thus case 24, plotted as X; and cases 11, 17, and 18, plotted as K, Q, and R.

To keep the discussion short, we will skip over x_1 and x_2. The image for x_3 shows a circular cluster for the 22 "clean" points. Superimposed on this is a downward-sloping trio of points K, R, and Q, and a remote horizontal point X. This clarifies the almost significant regression coefficient for x_3 completely: the significance is based entirely on case 24, and is reduced by the outlying cases 11, 17, and 18. The added variable plot however gives a different impression—here the 22 "clean" points also show an apparent linear relationship, providing a wrong diagnosis of the data. Thus the image plot shows the true nature of the relationship between Y and x_3 and its connection to the four aberrant cases, while the added variable plot is misleading.

It is not customary to produce added variable plots for the intercept, and so this has not been done. The image plot shows that with case 24 in the data

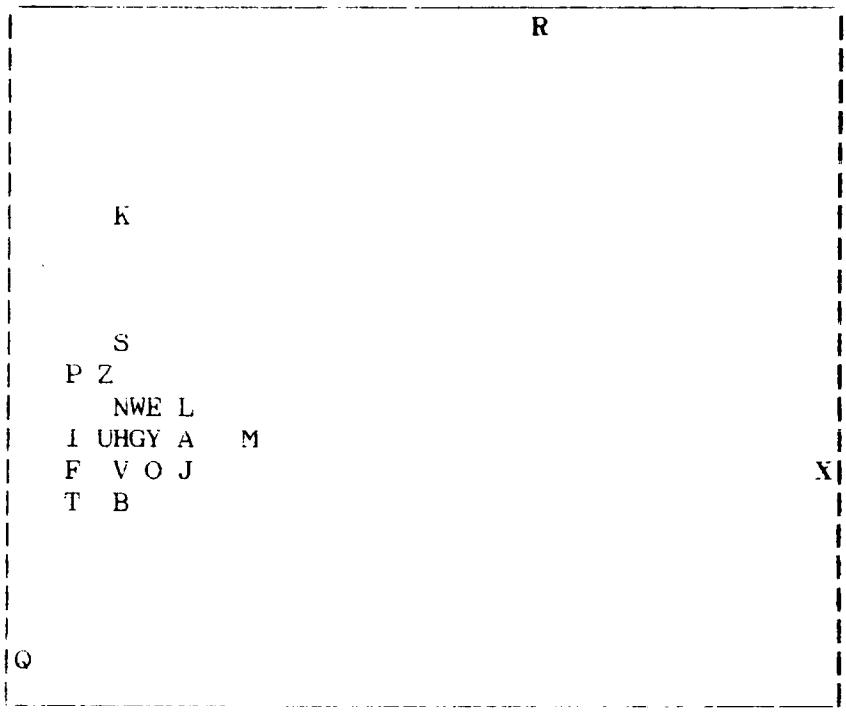

FIG. 3. Scatterplot of Y^* against x_3^*.

set, the intercept is very important, though when it is excluded, whether the trio of outliers is present or absent, there is no need for an intercept.

Thus the major features of this data set have been detected by a single pass of the MPIT analysis.

SUMMARY

The image or dual variates discussed in this paper have a number of valuable properties whose utility does not appear to be as widely known as it should. The dual matrix is particularly helpful in providing regression case diagnostics, and in the property that it generates the inverse covariance matrix of the dependent and predictor variables, and hence the statistics necessary for fitting multiple regressions. This connection extends to subset regression, where the elimination of one or more predictors from the primal regression can be studied by the inclusion of the corresponding variables in the dual regression.

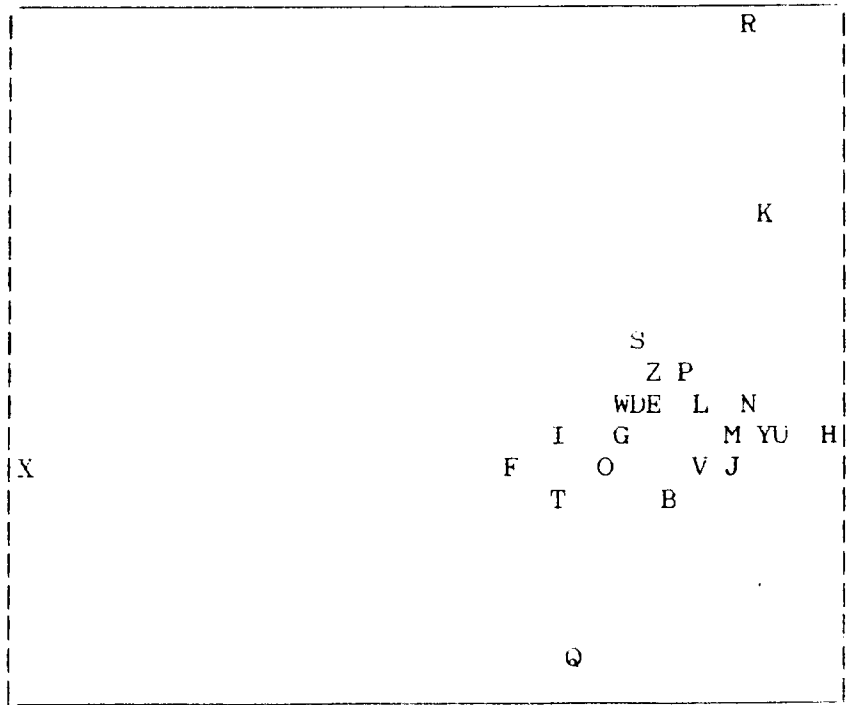

FIG. 4. Scatterplot of Y^* against x_4^*.

It might seem that the dual is defective in treating the predictor variables no differently from the dependent variable, but we argue that on the contrary this feature is valuable. Not only is nothing lost (since it is easy to compute all the standard regression statistics and case diagnostics from the dual matrix), but there is a positive benefit in that it allows one to explore the connections between the two sets of variables in a more symmetric way, and to uncover features of the data (such as the identification of outliers with possible errors in the predictors) that are not so easily obtained using the more traditional distinction between predictors and dependent variable. We thus suggest that the dual matrix is a powerful summary of all multiple regression information —both from predictors to dependent and vice versa—that warrants use as a standard case diagnostic.

The authors are very grateful to a referee for penetrating comments on the initial version of the paper, and a number of helpful suggestions for improvements.

REFERENCES

Belsley, D. A., Kuh, E., and Welsch, R. E. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York.

Dempster, A. P. 1969. *Elements of Continuous Multivariate Analysis*, Addison-Wesley, Reading, Mass.

Eplett, W. J. R. 1978. A note about the multipliers in latent root regression, *J. Roy. Statist. Soc. Ser. B* 40:184–185.

Gnanadesikan, R. and Kettenring, J. R. 1972. Robust estimates, residuals and outlier detection with multiresponse data, *Biometrics* 28:81–124.

Gragg, W. B., LeVeque, R. J., and Trangenstein, J. A. 1979. Numerically stable methods for updating regressions, *J. Amer. Statist. Assoc.* 74:161–168.

Gray, J. B. and Ling, R. F. 1984. *K* clustering as a detection tool for influential subsets in regression, *Technometrics* 26:305–320.

Graybill, F. 1969. *Introduction to Matrices with Applications in Statistics*, Wadsworth, Belmont, Calif.

Guttman, L. 1953. Image theory for the structure of quantitative variates, *Psychometrika* 18:277–296.

Haber, M. 1975. The Singular Value Decomposition of Random Matrices, Ph.D. Thesis, Hebrew Univ., Jerusalem.

Hawkins, D. M. 1973. On the investigation of alternative regressions by principal component analysis, *Appl. Statist.* 22:275–286.

Hawkins, D. M. 1974. The detection of errors in multivariate data using principal components, *J. Amer. Statist. Assoc.* 69:340–344.

Hawkins, D. M. and Eplett, W. J. R. 1982. The Cholesky factorization of the inverse covariance or correlation matrix in multiple regression, *Technometrics* 24:191–198.

Hocking, R. R. and Pendleton, O. J. 1983. The regression dilemma, *Comm. Statist. A* 12:497–527.

Kennedy, W. and Gentle, J. 1980. *Statistical Computing*, Marcel Dekker, New York.

Mosteller, F. and Tukey, J. W. 1977. *Data Analysis and Regression*, Addison-Wesley, Reading, Mass.

Pringle, R. M. and Rayner, A. A. 1971. *Generalized Inverse Matrices with Applications to Statistics*, Griffin, London.

Rao, C. R. and Mitra, S. K. 1971. *Generalized Inverse of Matrices and its Applications*, Wiley, New York.

Styan, G. P. H. 1983. Generalized inverses, in *Encyclopedia of Statistical Sciences* (S. Kotz, N. L. Johnson, and C. B. Read, Eds.), Wiley, New York.

Velleman, P. F. and Welsch, R. E. 1981. Efficient computing of regression diagnostics, *Amer. Statist.* 35:234–242.

Webster, J. T., Gunst, R. F. Mason, R. L. 1974. Latent root regression analysis, *Technometrics* 16:513–522.

Received 5 March 1988; final manuscript accepted 15 December 1988